Identifying Variables Responsible for Clustering in Discriminant Analysis of Data from Infrared Microspectroscopy of a Biological Sample

Martin, Francis L.; German, Matthew J.; Wit, Ernst; Fearn, Thomas; Ragavan, Narasimhan; Pollock, Hubert M.

Published in:
Journal of Computational Biology

JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 14, Number 9, 2007

Pp. 1176–1184
DOI: 10.1089/cmb.2007.0057

Identifying Variables Responsible for Clustering in Discriminant Analysis of Data from Infrared Microspectroscopy of a Biological Sample

FRANCIS L. MARTIN,[1] MATTHEW J. GERMAN,[2] ERNST WIT,[3] THOMAS FEARN,[4] NARASIMHAN RAGAVAN,[5] and HUBERT M. POLLOCK[6]

ABSTRACT

In the biomedical field, infrared (IR) spectroscopic studies can involve the processing of data derived from many samples, divided into classes such as category of tissue (e.g., normal or cancerous) or patient identity. We require reliable methods to identify the class-specific information on which of the wavenumbers, representing various molecular groups, are responsible for observed class groupings. Employing a prostate tissue sample divided into three regions (transition zone, peripheral zone, and adjacent adenocarcinoma), and interrogated using synchrotron Fourier-transform IR microspectroscopy, we compared two statistical methods: (a) a new "cluster vector" version of principal component analysis (PCA) in which the dimensions of the dataset are reduced, followed by linear discriminant analysis (LDA) to reveal clusters, through each of which a vector is constructed that identifies the contributory wavenumbers; and (b) stepwise LDA, which exploits the fact that spectral peaks which identify certain chemical bonds extend over several wavenumbers, and which following classification via either one or two wavenumbers, checks whether the resulting predictions are stable across a range of nearby wavenumbers. Stepwise LDA is the simpler of the two methods; the cluster vector approach can indicate which of the different classes of spectra exhibit the significant differences in signal seen at the "prominent" wavenumbers identified. In situations where IR spectra are found to separate into classes, the excellent agreement between the two quite different methods points to what will prove to be a new and reliable approach to establishing which molecular groups are responsible for such separation.

Key words: adenocarcinoma, biomedical, clustering, LDA, microspectroscopy, misclassification.

[1]Biomedical Sciences Unit, Lancaster University, Lancaster, United Kingdom.
[2]School of Dental Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom.
[3]Statistical Bioinformatics Group, Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom.
[4]Department of Statistical Science, University College London, London, United Kingdom.
[5]Lancashire Teaching Hospitals NHS Trust, Preston, United Kingdom.
[6]Department of Physics, University of Lancaster, Lancaster, United Kingdom.

1. INTRODUCTION

PROBLEMS MAY ARISE when groupings are to be identified from patterns obtained by the processing of high-dimensional data, such as those derived from microspectroscopy analysis using Fourier transform infrared (FTIR) methods. Such data are obtained from infrared (IR) spectral measurements on many samples where the variables (such as wavenumbers in spectroscopic data) may number several thousand. We will be concerned with multivariate pattern recognition ("cluster analysis"), rather than classification modelling where a training or calibration set of samples is available. Once any groupings (clusters) have appeared, class-specific information on which variables give rise to the observed separation of IR spectra into clusters is often required. Thus, one may infer the physical meaning of a factor by observing which original variables "load" most heavily onto it.

High-dimensional prediction problems are intrinsically difficult and beyond current techniques. One approach to make progress is to try and use more information implicitly known about the data. The issue of over-fitting is a perennial threat to be guarded against. There is a serious problem in applying linear discriminant analysis (LDA) directly to high-dimensional data such as IR spectra, just as there is in fitting regression equations with very large numbers of variables (Fearn, 2002). The large number of dimensions allows too much scope for discrimination to be achieved by chance, in directions that represent mainly noise. One way of reducing the dimensionality problem is simply to reduce the dimensionality of the feature space in a way that takes no account of the class labels. Principal component analysis (PCA) or related methods are ideal candidates, but the problem with these dimension reduction techniques is the lack of interpretability of the resulting classifiers.

Our aim was to find a means of obtaining class-specific information on which wavenumbers, and hence chemical bonds, are responsible for the observed grouping into clusters of IR spectra derived from a biological sample (prostate tissue). To this end, we compared two methods:

- The first uses a new variation of PCA that we term the "PCA-LDA cluster vector approach." The dimensions of the dataset are reduced, and LDA is then used to reveal clusters, through each of which a vector is constructed that identifies the contributory wavenumbers.
- The second method uses another way of reducing the dimensions of the same data set, namely stepwise LDA. It exploits the fact that the feature space in this problem is highly structured, in that measurements along the wavenumbers are strongly spatially correlated. In fact, it is well known that IR spectral peaks that identify certain chemical bonds extend over several wavenumbers. Following classification via either one or two wavenumbers, the method checks whether the resulting predictions are stable across a range of nearby wavenumbers.

2. METHODS

We employed a set of IR spectroscopic data derived from a previous publication (German et al., 2006), in particular those data whose analysis is summarized in Figure 5 of that paper.

2.1. Biological samples

As described in German et al. (2006), informed consent to obtain a prostate tissue set for research was obtained (LREC no. 2003.6.v; Preston, Chorley and South Ribble Ethical Committee) from patients ($n = 6$) undergoing radical retropubic prostatectomy for prostate adenocarcinoma (CaP). A CaP-free prostate tissue mass separate to the area of CaP, which would be present in the gland, was isolated post-surgery. Tissue slices representing the peripheral zone (PZ) or transition zone (TZ) were dissected out from these tissue masses. Independently of this procedure, a slice containing CaP and isolated from a different part of the prostate gland was also retrieved. Subsequently, a microtomed 10-μm-thick section of each tissue slice was prepared for analysis by IR spectroscopy. These microtomed sections were floated onto 0.5-mm-thick BaF_2 windows (Photox Optical Systems, Sheffield, UK) for transmission-mode synchrotron FTIR microspectroscopy. The sections were then dewaxed by immersion in fresh xylene (5 min) and then washed in absolute alcohol (74OP; 5 min), to remove the xylene. Spectra were acquired at Daresbury (Warrington, UK) synchrotron source on beamline 11.1 using the Thermo Nicolet Continuum microscope

and Nexus FTIR spectrometer. Spectral collection was in transmission mode, and spectra were converted to absorbance using Thermo Omnic software. A ×32 Reflachromat objective was used, and the aperture area was 10 × 10 μm. Spectra were collected at 4 cm^{-1} spectral resolution and co-added for 1024 scans. In CaP-free tissues (PZ or TZ), five independent spectral measurements were taken on each of three randomly chosen glandular elements. In CaP regions, three independent spectral measurements were taken on each of five randomly chosen glandular elements. Spectra were baseline corrected using OPUS software and normalized to the amide II (1533 cm^{-1}) absorbance band.

2.2. The PCA cluster vector approach

IR microspectroscopic studies of biological samples can involve the processing of spectral measurements derived from many samples, divided into classes such as category of sample (Walsh et al., 2007) or patient identity (German et al., 2006). Interpretation typically yields two types of information: clustering and the identification of the chief contributory variables. One may group spectra into clusters and determine the extent to which these clusters correspond to classes of sample, using PCA, which is built on the assumption that variation implies information. It replaces the original several hundred wavenumber variables by linear combinations thereof, termed principal components (PCs), which seek to capture as much variability as possible. The PCs are automatically listed in order, according to how much of the original data variance is accounted for by each one: often, typically no more than the first 10 or so need to be taken into account. For each spectrum obtained, the whole set of many hundreds of readings (one for each wavenumber) is replaced by a small number of "scores," one for each significant PC. Thus in a scores plot, each set of measurements (IR spectrum) appears as a single point, whose coordinates are its scores on the one, two or more PCs chosen as axes for the plot (Fig. 1).

In PCA, plots showing the extent to which each wavenumber contributes to a given PC constitute a pseudo-spectrum or "loadings plot," giving the class-specific information on which of the spectral peaks (and molecular constituents) are responsible for the observed groupings. Comparing line plots of the original data and loadings lets one see which data features are captured by a particular PC. However, in general no single PC will pass through the median of any of the particular clusters of interest. Hence, we decided to address the need to construct a "cluster vector," which will likewise exhibit a loadings plot, even though it will not be a true PC in that it is not in general orthogonal to all the others (German et al., 2006).

Previously, in order to derive a loadings plot (pseudo-spectrum) to define those regions of the spectra that exhibit bio-molecular and/or conformational changes, we developed a semi-graphical method that constructs the cluster vector, passing through the median of the cluster (see the Supplementary Data in German et al. [2006]). Its loadings plot may be calculated simply by taking a weighted average of the components of the loadings of three real PCs that were used to identify the cluster. Thus, the method was a graphical version of the following procedure:

a. In PCA, identify the three PCs which, when used as axes in hyperspace for the cluster plot, allow it to be rotated to the point where its projection gives the best cluster separation (Fig. 1A).
b. For each of these PCs (e.g., nos. 1, 2, 4), find the median score for the samples in the cluster that one wants to characterize.
c. Sum the three loadings vectors (given by the PCA software) for PCs 1, 2, 4 weighted by the median scores.
d. The resultant is a loadings vector that is plotted as an effective loadings plot for that cluster (Fig. 1B), even though it is not orthogonal to the PCs themselves.

A potential disadvantage of using PCA alone is that it does not unambiguously give the optimum grouping into clusters. A similar approach exploits the benefit of linear discriminant analysis (or canonical variants analysis, LDA), using PCA for preliminary dimensionality reduction (Fearn, 2002; Walsh et al., 2007). LDA explicitly attempts to model the differences between the classes of data that were assigned *a priori*. New variables (linear discriminants [LD]) are found such that the ratio of the between-cluster variance to the within-cluster variance is maximized, and thus the clusters appear seen at maximum separation. Thus LDA, like regression methods such as Partial Least Squares, is a "supervised" method, in that it requires some knowledge (classes) of the spectra of the sample's constituents, as indeed was

necessary in (b) above. Consider N samples which separate into h clusters, the number of original variables (wavenumbers) being m: the number of significant PCs used in the analysis is g, and the subsequent LDA analysis yields d LD coefficients as outlined below. In this case we have 45 samples, 3 clusters, 234 wavenumbers, 7 PCs, and 2 LDs. The following notation employs the subscripts i, j, k, w to refer to individual samples, PCs, LD coefficients, and wavenumbers, respectively. PCA gives a 45×7 scores matrix $\mathbf{S}_{Ng}$, a 234×7 loadings matrix $\mathbf{L}_{mg}$, and a 3-element column vector of sample classes $\mathbf{c}_N$. LDA is then used to process data in the form of $\mathbf{c}_N$ and $\mathbf{S}_{Ng}$, to give a 2×7 LD coefficients matrix $\mathbf{D}_{dg}$. Using these LD coefficients as weighting factors for the PCA scores, the 45×2 LD scores matrix $\mathbf{SD}_{ikh} = \sum(\mathbf{D}_{kj}\mathbf{S}_{ij})$ is then computed. This gives the LDA scores plot.

LDA has the additional advantage that it allows a choice of the predetermined classes (in this case, type of tissue or identity of patient) to be taken into account during the derivation of loadings plots as well as clusters. The PCA-LDA cluster vector procedure is as follows: for each of the h clusters, from $\mathbf{SD}_{ikh}$ the 2×3 mean cluster scores matrix $\mathbf{mD}_{kh}$ is calculated, then the 234×2 LD loadings matrix $\mathbf{LD}_{wk} = \sum_j(\mathbf{D}_{kj}\mathbf{L}_{wj})$, and finally the 234×3 cluster loadings matrix $\mathbf{E}_{mh} = \sum_h\sum_k(\mathbf{LD}_{wk}\mathbf{mD}_{kh})$. Plotting a single column (fixed value of h) then gives the loadings plot for an individual cluster.

2.3. The "stepwise LDA" method

As mentioned previously, here we consider classification via either one or two wavenumbers, and then check whether the resulting class predictions are stable across a range of nearby wavenumbers (Wit and McClure, 2004). The prediction may be linear, as used here, or quadratic. The steps are as follows:

a. Any single prominent spectral peak is chosen, and LDA is used to process the data for all IR spectra at this wavenumber. It finds a new axis in the usual way, by maximizing the ratio of between-group variance to within-group variance. Thus the best Gaussian "posterior density" is fitted to each group.
b. Then one data point can be left out and the Gaussian is used to calculate the "predicted posterior class probability" for that spectrum, across the h class categories, at the one wavenumber. This is repeated for all IR spectra, omitting one data point at a time.
c. The above is then repeated for all wavenumbers.
d. For each data point, each of the predicted class probabilities is then compared with the known true value (either 1 or 0).
e. For each wavenumber, these absolute deviations are now summed to give a sum of absolute posterior misclassification or "misclassification error." Troughs on a plot of this error against wavenumber are then analogous to peaks on a PCA loadings plot, in that they indicate wavenumbers that best distinguish one particular class from the rest.

If this single wavenumber predictor approach yields a relatively high number of misclassifications, it is worthwhile to consider two or more wavenumbers simultaneously. It could be tempting to consider taking the wavenumbers having the lowest misclassification errors as shown by the single-predictor method: however, in case the two wavenumbers contain very similar information, in step 3 above one considers all possible pairs of wavenumbers. For each pair, the best Gaussian mountain is fitted to each group of data, and a three-dimensional graph of misclassification error as a function of wavenumbers 1 and 2 is plotted.

3. RESULTS

3.1. PCA-LDA

The prostate tissue sample included three regions: TZ, PZ, and adjacent CaP. A previously-published analysis of the data (German et al., 2006) was performed by means of PCA alone, and for comparison, some of those earlier results are shown in Figure 1A, showing the separation of data from the three regions into three distinct clusters, and Figure 1B, the loadings plot. When the new method is applied to the same raw data, as shown in Figure 1C, we obtain a PCA-LDA scores plot; Figure 1D shows the corresponding loadings plot for the CaP cluster. When compared with the results obtained by PCA alone, there are discrepancies, which may arise from the fact that the procedures described so far give loadings for vectors

that point from the origin to the centers of each cluster. This origin is somewhat arbitrary, and would shift somewhat if the numbers of samples in the clusters changed. There would be a major shift if, for example, a fourth cluster were introduced.

Accordingly, it seems clear that these methods are likely to be at their most reliable when data giving just two clusters are processed at any one time. An example is shown in Figure 2. Here, using the new PCA-LDA method, we have shown data from six patients (German et al., 2006). There must be many other factors (e.g., age, diet, disease progression) that contribute to spectral variations between different patients, and we have also examined the variation between patients, using data from a single tissue type and allocating a different class variable to each individual's samples. As expected, inter-patient variations are often significant, yet despite this, this method succeeds in picking out at least four prominent wavenumbers as being responsible for the clustering of spectra according to the relevant region of prostate tissue (PZ, TZ, or adjacent CaP).

3.2. Stepwise LDA

As explained above, the method involves classification via either one wavenumber (the single linear predictor version) or two wavenumbers. A check of the extent to which the resulting predictions are stable across a range of nearby wavenumbers then yields an absolute posterior misclassification or "misclassification error." Alternatively, misclassification error may be calculated and plotted as a function of two wavenumbers. Figure 3 shows a plot of misclassification error against wavenumber, as derived when the single linear predictor version is applied to the same raw data as before. The troughs are analogous to peaks on a PCA loadings plot, indicating the wavenumbers that best distinguish one particular class from the rest. Figure 4 shows a three-dimensional graph of misclassification error as a function of two wavenumbers. In this case it is observed, for example, that the two wavenumbers of approximately 1070 and 1135 cm^{-1} are prominent. Thus, differences in signal at these wavenumbers should provide a useful identification of certain molecular groups that occur in different concentrations between the different classes (i.e., regions) of prostate tissue (PZ, TZ, or adjacent CaP). For comparison with the loadings plots given by the PCA-LDA method, it is useful to convert data such as those of Figure 4A into a line plot of average misclassification error (i.e., the sum of absolute posterior misclassification); this is shown in Figure 4B.

4. DISCUSSION

We have presented the results of two methods for identifying the variables responsible for cluster formation: a variation of PCA-LDA that we term the cluster vector approach, and another way of reducing the dimensions of the same data set, known as the stepwise LDA method. Both methods were applied to the same sets of previously published data (German et al., 2006). In this case, when comparing Figures 3 and 4B, we see very little difference between the misclassification results given by the two variants of the stepwise LDA method (single predictor vs. two predictor).

FIG. 1. Processed spectral data acquired from epithelial cells lining glandular elements of prostate tissue from one patient, acquired using synchrotron Fourier transform infrared (FTIR) microspectroscopy. The spectra were collected from particular tissue regions (peripheral zone, black circles; transition zone, blue squares; and adjacent adenocarcinoma, red triangles). **(A)** Three-dimensional scores plot on PCs 1, 2, and 4, selected to demonstrate best cluster segregation using PCA (from Fig. 5A of German et al. [2006]). **(B)** Loadings plot corresponding to the cluster vector for the "red" cluster, shown dotted in A (from Fig. 4D of German et al. [2006]). **(C)** The same data as in A, processed using PCA-LDA. **(D)** The same data as in B, processed using PCA-LDA.

FIG. 2. Data obtained from six patients and processed using PCA-LDA, two tissue types and one patient at a time (transition zone [TZ], adjacent adenocarcinoma [CaP]; compare part of Fig. 4D of German et al. [2006]). Other details are as in Figure 1. **(A)** Scores plots, showing complete cluster separation. **(B)** Loadings plots. **(C)** As in B, showing detail of the low-wavenumber region.

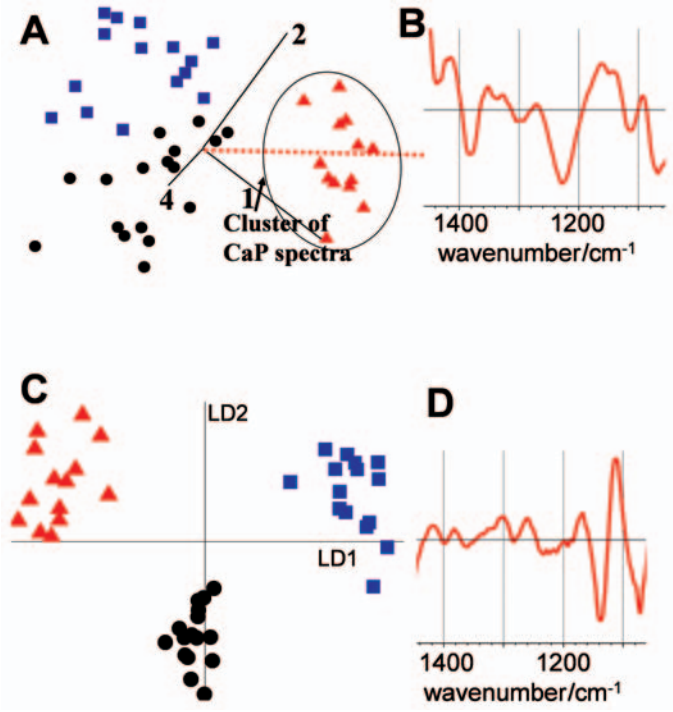

FIG. 1. (*See caption on page 1180.*)

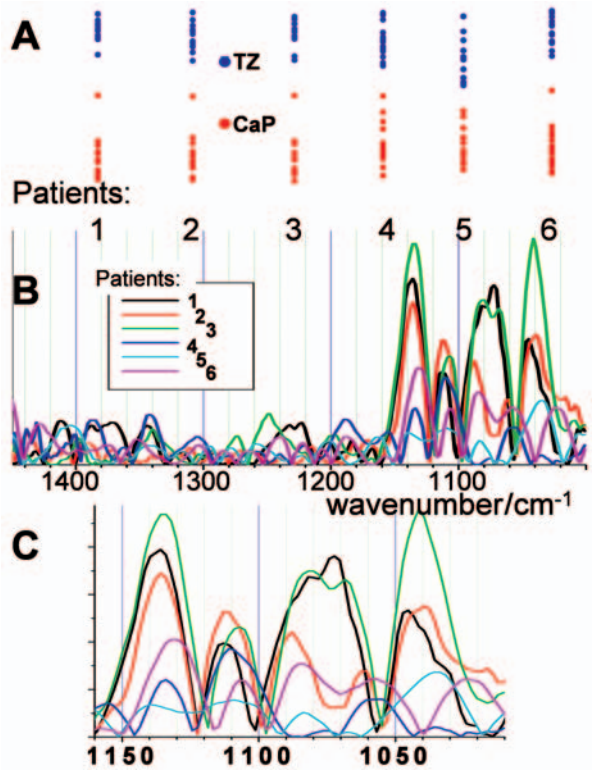

FIG. 2. (*See caption on page 1180.*)

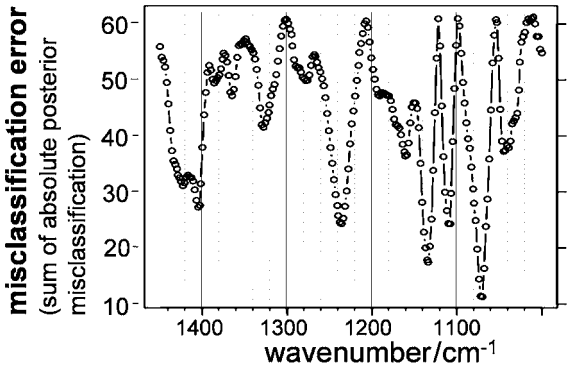

FIG. 3. The same raw data as in Figure 1, processed using the single linear predictor stepwise LDA method and showing misclassification error as a function of wavenumber. Significant wavenumbers are shown by troughs.

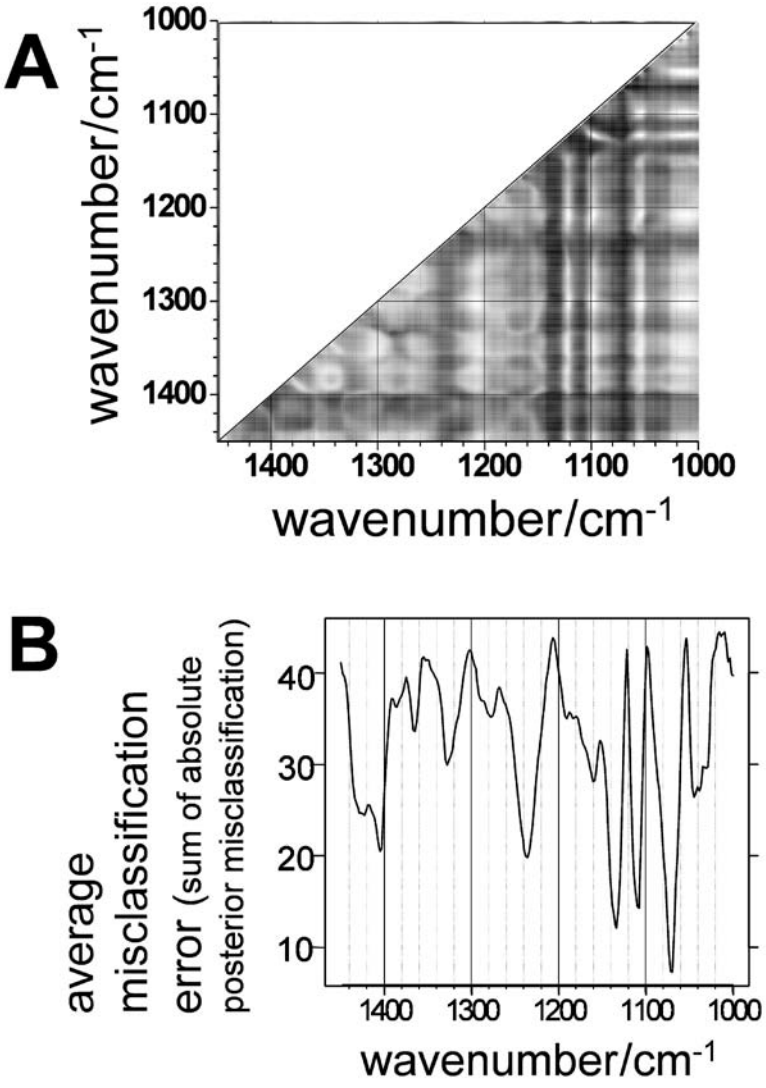

FIG. 4. **(A)** As in Figure 3, processed using the two-wavenumber predictor method. Black indicates low misclassification errors and good prediction values. **(B)** The data of Figure 4A re-plotted to show average misclassification error (i.e., the sum of absolute posterior misclassification). Troughs show significant wavenumbers.

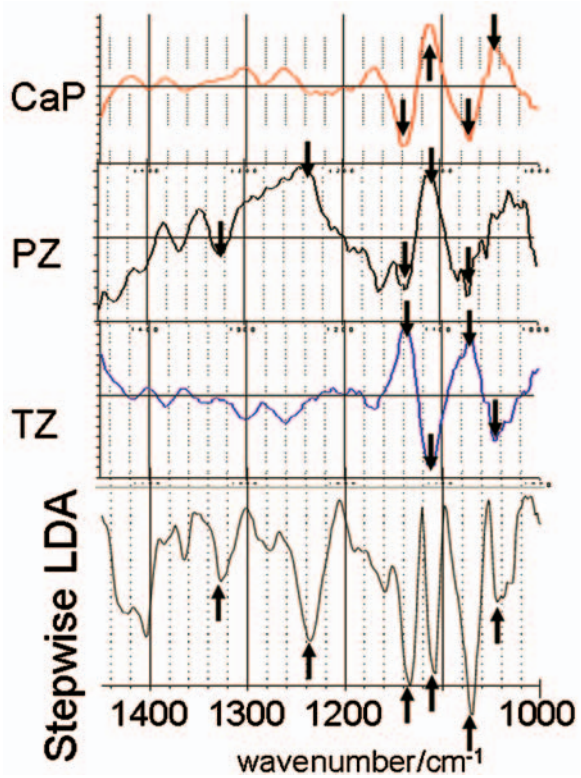

FIG. 5. Comparison of the plots shown by PCA-LDA and stepwise LDA, for the same raw data sets: as in Figure 1D, but showing loadings plots for all three clusters, versus result of two-wavenumber predictor method (stepwise LDA; Fig. 4B). Significant agreement is shown by arrows. These appear as troughs in the stepwise LDA plot, and as either peaks or troughs in the PCA-LDA plots.

In Figure 5, the predictions of PCA-LDA and stepwise LDA are compared, using the one-patient data as an example:

a. As shown in Figure 1D, the PCA-LDA wavenumbers show four very prominent (positive or negative) peaks for the CaP cluster. Also shown in Figure 5 are five peaks (or troughs) for the PZ cluster, and three for the TZ cluster, which derive from the relevant PZ and TZ data. One needs to keep in mind that the three PCA-LDA pseudo-spectra shown are linear combinations of just two LDA loadings, so it is not surprising that one sees the same pattern repeated in all three, or that the signs change (see the comment on the origin of the loading vectors, 3.1 above). In a situation where there are only two clusters and hence one LDA loading, the two plots would be mirror images of each other. We see that *all* these wavenumbers closely coincide with the predictions of the stepwise LDA method.
b. The troughs indicated by stepwise LDA are all confirmed by PCA-LDA, with one exception (at 1405 cm^{-1}).

For example, as regards the molecular groups that distinguish the adjacent adenocarcinoma (CaP), both methods pick out the wavenumbers 1135 cm^{-1} (C−O ring vibrations of nucleic acid "sugars"), 1110 cm^{-1} (phosphate species such as adsorbed H_2PO_4−, also C=O stretching and bending of ketones, C−H in-plane vibrations of polyimide aromatic rings), 1075 cm^{-1} (PO_2− symmetric stretching vibrations of nucleic acids and phospholipids), and 1050 cm^{-1} (ribose-phosphate main-chain vibrations). This result is consistent with the conclusions of our previously-published findings (German et al., 2006), using the same raw data; in addition, it clarifies some of the uncertainties detailed there.

The remarkable degree of agreement between the two methods, the PCA-LDA cluster vector approach and stepwise LDA, is encouraging. Stepwise LDA has the advantage of simplicity. For example, in appropriate cases it will be very convenient to use just one of the pairs of wavenumbers indicated, in order to pick out the key molecular groups responsible for clustering. However, stepwise LDA does not, in its

present form, indicate which of the different classes of spectra are exhibiting the significant differences in signal that are seen at the prominent wavenumbers identified. This is not necessarily a major disadvantage, since as already remarked in connection with PCA-LDA, the identification of the variables responsible for cluster formation is likely to be most reliable when just two classes are processed at any one time. In such a case, there is no ambiguity about which class is involved. Accordingly, we claim that in situations where IR spectra are found to separate into classes, this study points to what could prove to be a new and reliable approach to establishing which molecular groups are responsible for such separation.

ACKNOWLEDGMENTS

We thank Dr. J.M. Chalmers and Dr. A.M.C. Davies for valuable suggestions. This work was performed as part of a project supported by funding from EPSRC and Rosemere Cancer Foundation.

REFERENCES

Fearn, T. 2002. Discriminant analysis, 2086–2093. In: Chalmers, J.M., and Griffiths, P.R., eds., *Handbook of Vibrational Spectroscopy. Vol. 3*. Wiley, New York.

German, M.J., Hammiche, A., Ragavan, N., et al. 2006. Infrared spectroscopy with multivariate analysis potentially facilitates the segregation of different types of prostate cell. *Biophys. J.* 90, 3783–3795.

Walsh, M.J., Singh, M., Pollock, H.M., et al. 2007. ATR microspectroscopy with multivariate analysis segregates grades of exfoliative cervical cytology. *Biochem. Biophys. Res. Commun.* 352, 213–219.

Wit, E., and McClure, J. 2004. *Statistics for Microarrays*. Wiley, Chichester, UK.

Address reprint requests to:
Dr. Hubert M. Pollock
Department of Physics
University of Lancaster
Lancaster LA1 4YB, UK

E-mail: h.m-pollock@lancaster.ac.uk